Fast, High-Quality Hierarchical Depth-Map Super-Resolution

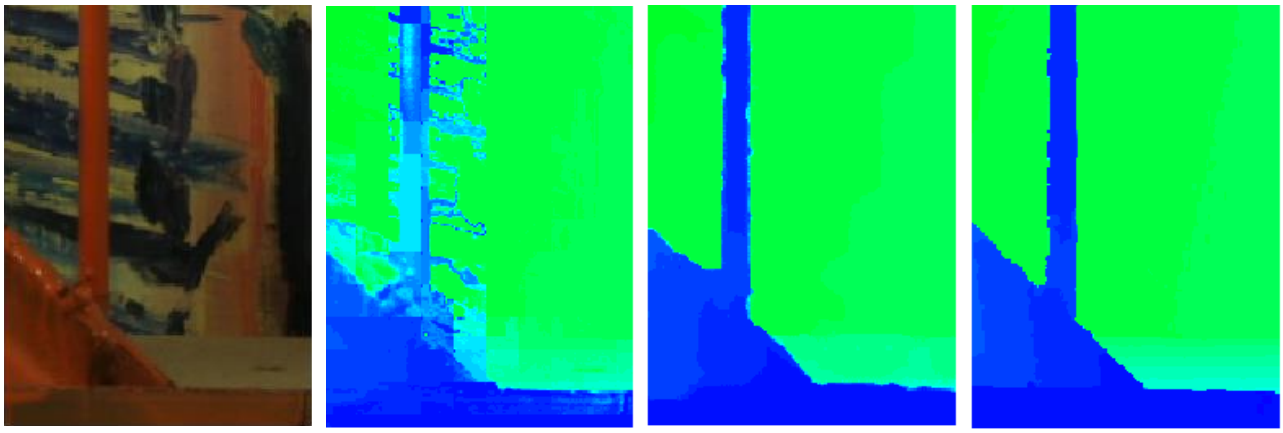
Yiguo Qiao
University of Bath
Bath, UK
yiguo.qiao@bath.edu

Licheng Jiao
Xidian University
Xi'an, China
lchjiao@mail.xidian.edu.cn

Wenbin Li
University of Bath
Bath, UK
wenbin.li@bath.edu

Christian Richardt
University of Bath
Bath, UK
christian@richardt.name

Darren Cosker
University of Bath
Bath, UK
D.P.Cosker@bath.ac.uk



(a). RGB image (b). JBU method (c). Our method (d). Ground truth

Figure 1: Our proposed method (C-HDS) generates 16× high-resolution depth maps with clear and sharp edges.

ABSTRACT

The low spatial resolution of acquired depth maps is a major drawback of most RGBD sensors. However, there are many scenarios in which fast acquisition of high-resolution and high-quality depth maps would be desirable. One approach to achieve higher quality depth maps is through super-resolution. However, edge preservation is challenging, and artifacts such as depth confusion and blurring are easily introduced near boundaries. In view of this, we propose a method for fast, high-quality hierarchical depth-map super-resolution (HDS). In our method, a high-resolution RGB image is degraded layer by layer to guide the bilateral filtering of the depth map. To improve the upsampled depth map quality, we construct a feature-based bilateral filter (FBF) for the interpolation, by using the extracted RGB shallow and multi-layer features. To accelerate the process, we perform filtering only near depth boundaries and through matrix operations. We also propose an extension of our HDS model to a Classification-based Hierarchical Depth-map Super-resolution (C-HDS) model, where a context-aware trilateral filter reduces the contributions of unreliable neighbors to the current missing depth location. Experimental results show that the proposed method is significantly faster than existing methods for generating high-resolution depth maps, while also significantly improving depth quality compared to the current state-of-the-art approaches, especially for large-scale 16× super-resolution.

MM '21, October 20–24, 2021, Virtual Event, China

ACM ISBN 978-1-4503-8651-7/21/10…$15.00
https://doi.org/10.1145/3474085.3475595

CCS CONCEPTS

• **Computing methodologies** → **Image processing**; **Computational photography**; • **Mathematics of computing** → *Interpolation.*

KEYWORDS

hierarchical depth-map super-resolution, edge preservation, feature-based bilateral filter, context-aware trilateral filter

ACM Reference Format:
Yiguo Qiao, Licheng Jiao, Wenbin Li, Christian Richardt, and Darren Cosker. 2021. Fast, High-Quality Hierarchical Depth-Map Super-Resolution. In *Proceedings of the 29th ACM International Conference on Multimedia (MM '21), October 20–24, 2021, Virtual Event, China.* ACM, New York, NY, USA, 10 pages.
https://doi.org/10.1145/3474085.3475595

1 INTRODUCTION

Stereo vision has applications across many areas, including robot navigation, 3D measurement and virtual reality. Depth acquisition – the core aspect of stereo vision – has received widespread attention from researchers [1, 27, 29, 44, 46]. Active sensing techniques, like laser range finders, are popular ways to obtain depth information [43, 45]. Microsoft's Kinect sensor [14, 37, 48], based on time-of-flight (ToF), is one such range finder [10, 20]. Both a regular RGB camera and a depth sensor are incorporated into the Kinect. However, despite their popularity, the low-resolution depth maps obtained from RGB sensors is a drawback of present devices. To alleviate this issue, the ability to generate high-resolution depth maps is an attractive solution – with the corresponding RGB image (captured at a far higher resolution) being an ideal guide for generating super-resolution versions of the degraded depth map [3, 5, 24, 33, 42].

The focus of depth super-resolution is to exploit the relationship between the RGB image and the depth map as much as possible, allowing the RGB colors to guide upsampling of the depth. However, this relationship is not straightforward for the following two reasons. First, the spatial distances between the points/pixels to be predicted and the guiding points are determined by the magnification (or scaling) factor. The higher the magnification, the larger the distance and the harder it is for the relationship between RGB and depth values to be accurately exploited. Second, the relationship between the RGB image and the depth map is neither linear nor stable across the modalities, varying from region to region. This relationship may lead to unexpected artifacts, such as depth blurring, depth confusion, depth bleeding and missing depth values – especially in boundary areas.

To address these issues, we present a novel hierarchical depth-map super-resolution method (HDS), in which a high-resolution RGB image is degraded layer by layer to guide the bilateral filtering of the depth map. Moreover, a classification-based HDS (C-HDS) is proposed to reduce the contributions of unreliable neighbors to the current missing depth location, resulting in sharp, clear depth edges as shown in Figure 1. Our approach outperforms other state-of-the-art depth-map super-resolution methods in terms of reconstruction error and performance on established benchmarks, and has the following attributes and contributions:

- A hierarchical image pyramid is adopted to shorten the long spatial distances (especially in the case of large super-resolution scales) between the to-be-interpolated depth locations and their neighboring known-depth locations, through the layer-by-layer operation. This significantly reduces the interpolating errors.
- A boundary-aware interpolation and a parallel matrix operator are introduced to accelerate computation (5.93s v.s. 1.64s at 16× scale upsampling).
- Joint convolutional features, i.e. shallow features and multi-layer features, are extracted from the RGB image. The former mainly contributes to edge-preservation, while the latter is mainly used for structural integrity.
- A novel local depth guided RGB classification is applied to detect sharp depth edges hidden in RGB image. This detected depth information makes significance in edge preserving while generating upper scale depth.

The remainder of the paper is organised as follows. In Section 2, we briefly review the related work. In Section 3, we describe the proposed basic HDS. In Section 4, we provide the updated C-HDS. Experimental results are shown Section 5. We conclude the paper in Section 6.

2 RELATED WORK

Many approaches to super-resolution have previously been proposed [e.g. 28, 49, 50], broadly across three categories: filtering-based, optimization-based and learning-based.

Filtering-Based Depth-Map Super-Resolution. In most filtering-based methods, missing pixels are interpolated but existing depth pixels are left unchanged [21, 24, 25, 32]. Joint bilateral upsampling (JBU) [21] is based on a joint bilateral filter[8, 31], which is influenced by both color difference and spatial distances on depth map (e.g. see Figure 1b). JBU interpolates blank pixels from all neighboring non-blank pixels, while joint geodesic upsampling (JGU) [24] instead selects a set of the best non-blank pixel candidates using a global geodesic search. Joint trilateral filter based upsampling [25] first upsamples the low-resolution depth map with a simple interpolation method before refining the boundaries of the intermediate result in an outside-in order. Segmentation-based upsampling (SBU) [32] first converts the depth super-resolution problem into an RGB image segmentation problem. Based on this, a joint trilateral filter is constructed to locally interpolate the low-resolution depth.

Optimization-Based Depth-Map Super-Resolution. In these approaches, a data term is used to maintain depth consistency between the upsampled result and the initial low-resolution depth map. A regularization term is used to preserve the edges in the upsampled result, to coincide with the color image as much as possible. Different methods in this domain are generally distinguished by their regularization terms. Methods using Markov random fields (MRF) [7] regularize according to color difference. Total generalized variation (TGV) method [9] formulate regularization using an anisotropic diffusion tensor. Adaptive auto-regressive (AR) method [47] use an AR predictor as the regularization term. In the optimization-based static/dynamic (SD) filtering method [13], the high-resolution RGB image and the intermediate depth of the last optimization iteration are used as static guidance and dynamic guidance respectively.

Learning-Based Depth-Map Super-Resolution. With the success of deep neural networks in computer vision in recent years, some deep learning based methods have also been developed. In Li et al. [23], a joint filter is constructed based on CNNs to selectively transfer salient structures that are consistent in both guidance RGB image and target high-resolution depth maps. Hui et al. [17] proposes a deep learning method to take advantages of upsampling different spectral components, then further achieves super-resolution depth map. Their improved version i.e. Voynov et al. [42], measures the quality of depth-map super-resolution using renderings of resulting 3D surfaces. In addition, by comparing with a number of existing perceptual metrics, they also proved that a visual appearance based loss yields significantly improved 3D shapes. Deep Image Prior [41] is a learning structure to sufficiently capture low-level image statistics prior to any learning. Such prior is reported [42] to naturally and efficiently allow super-resolution for depth.

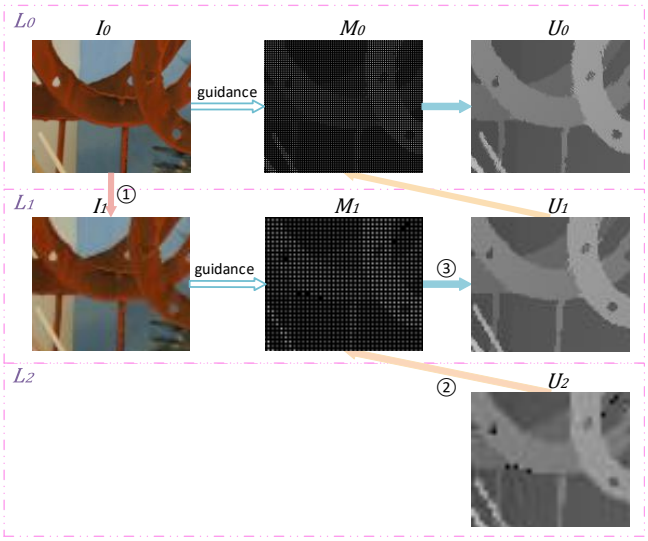


Figure 2: Overview of the proposed hierarchical depth-map super-resolution (HDS). Three steps are: (1) 2× RGB image downsampling, (2) mapping the low-resolution depth map to the high-resolution grid, and (3) depth map interpolation under the guidance of RGB image in the same layer.

The advantages and disadvantages of these methods are discussed in detail in Section 5.1, through comparison with the proposed method.

3 HIERARCHICAL DEPTH-MAP SUPER-RESOLUTION

In this section, we first give an overview of the our HDS method. We then improve the quality of the upsampled depth maps using RGB-based convolutional feature extraction. Finally, to improve efficiency, boundary detection and matrix operations are introduced.

3.1 Overview

Figure 2 shows an overview of our basic HDS method (under a super-resolution scale of 4×), which contains three steps [33, 39]. First, the RGB image I_0 is downsampled by a factor of 2× in each layer, to get the low-resolution RGB images $\{I_1, I_2, ...I_n\}$. Then, from the last layer (lowest resolution), we map the low-resolution depth map U_n to the high-resolution grid in the layer above to get the input depth map M_{n-1}. Finally, the sparse depth map M_{n-1} is interpolated by JBU [21] under the guidance of the RGB image I_{n-1} in the same layer.

JBU is conducted based on a joint bilateral filter that takes as input a low-resolution depth D and a high-resolution RGB image $\widetilde{I}$:

$$\widetilde{D}_p = \frac{1}{W_p} \sum_{\overline{q}\in\Omega, D_{\overline{q}}\neq 0} D_{\overline{q}} \cdot w_d(\overline{p},\overline{q}) \cdot w_c(\widetilde{I}_p, \widetilde{I}_q). \quad (1)$$

Here, $\widetilde{D}$ is the output high-resolution depth map, p and q are (high-resolution) pixel locations, $\overline{p}$ and $\overline{q}$ are the corresponding downsampled pixel locations, W_p is a normalizing factor, Ω is the set of all neighboring pixel locations in the low-resolution depth map D, and w_d and w_c are two Gaussian weights for spatial and color distance, respectively.

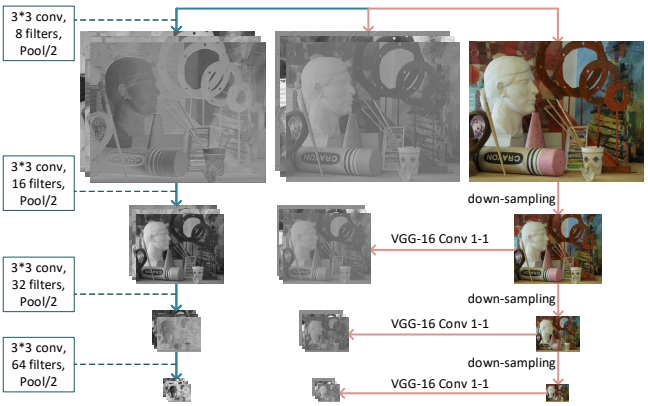


Figure 3: Convolutional feature extraction. The shallow feature extraction via Conv1-1 of VGG-16 is indicated by red lines while the multi-layer feature extraction based on the proposed multi-layer convolutional downsampling is indicated by blue lines.

Note that both the downsampling and the mapping are extracting or filling an image every other row or column so that correspondence between the same points of RGB image and depth map of the same layer is obtained. The total number of layers n can be calculated as $\log_2 r$, where r denotes the super-resolution scale.

3.2 Feature-Based Bilateral Filtering

RGB Image Feature Extraction. To improve the quality of the upsampled depth maps, a convolutional downsampling together with the VGG-16 network [38] is used to extract rich features, that is, the feature maps or context maps from the RGB image [30]. Deep neural networks (DNN) have been widely used in many areas due to the powerful feature extraction capabilities of the convolutional layers. Many classical DNN, like AlexNet [22], GoogleNet [40] and VGG-16 [38] perform well in general image feature extraction tasks. With the extracted context maps, a feature-based bilateral filter (FBF) is further constructed to achieve the feature based hierarchical depth map super-resolution (F-HDS) [11, 26].

Both shallow features and multi-layer features are extracted in feature extraction process. Specifically, we use the first layer of VGG-16 to extract the shallow features in RGB image. In addition, a convolutional downsampling is used to achieve the RGB image multi-layer feature exaction, in which random kernels are utilized in each layer. Figure 3 shows the framework of the proposed feature extraction under a 16× super-resolution scale. In practice, we use the first n layers in the case of smaller super-resolution scale.

For shallow feature extraction, we first downsample the RGB image by deleting every other row or column for each image layer. Then, the downsampled image is fed into the first convolutional layer of VGG-16 to obtain shallow feature maps. For multi-layer feature extraction, a convolutional downsampling strategy is proposed. We set the convolutional layer and the max-pooling layer parameters as shown in Figure 3. The number of the convolutional layers is determined by the super-resolution scale. The stride of the pooling layers is set to 2, which is suitable for the 2 times downsampling of the RGB image each layer. This convolutional downsampling preserves the edges while extracting the multi-layer features.

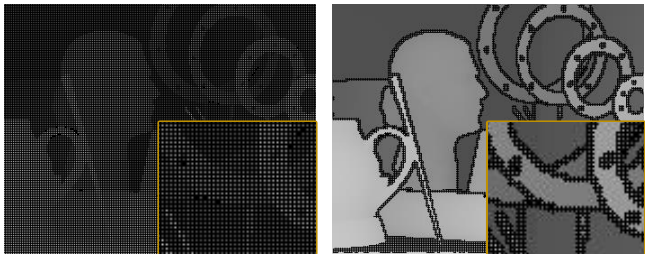


Figure 4: Visual results of depth boundary regions detection.

Feature-Based Bilateral Filter Construction: With the extracted feature maps, the joint bilateral filter described in Equation 1 can be promoted to feature-based bilateral filter:

$$\widetilde{D}_p = \frac{1}{W_p} \sum_{\overline{q}\in\Omega, D_{\overline{q}}\neq 0} \sum_k D_{\overline{q}} \cdot w_d(\overline{p},\overline{q}) \cdot w_c(\widetilde{F}_p^k, \widetilde{F}_q^k), \tag{2}$$

where F^k denotes the kth layer of the feature maps.

3.3 Performance Optimization

Bilateral filtering is generally performed pixel by pixel, leading to a large computational expense. Some previous work [4, 6] is able to speed up the filtering by approximating the final results while this may bring unexpected errors. We therefore propose a concise mechanism to accelerate the performance of our method in the following two ways.

Depth Boundary Detection. We first detect edges in the depth map so that bilateral filtering is performed only in these regions, see Figure 4. That is, we accelerate the program by reducing the number of interpolation operations. The edges can be simply detected by using a sliding window on an intermediate bilinear upsampled depth map. A pixel p will be treated as a boundary point if the depth range within this window is larger than a threshold, i.e.

$$\max_{q\in N(p)} D_q^* - \min_{q\in N(p)} D_q^* > \tau_b, \tag{3}$$

where D^* denotes the intermediate bilinear upsampled depth map, $N(p)$ denotes all neighbouring pixels of p, and $\tau_b = 5$ is the depth threshold we use. Then only the depth boundary regions are interpolated with JBU, the rest regions are filled with the corresponding regions in the intermediate bilinear upsampled depth map.

Matrix Computation. We first expand Equation 1:

$$\widetilde{D}_p = \frac{1}{W_p} \sum_{\overline{q}\in\Omega, D_{\overline{q}}\neq 0} D_{\overline{q}} \cdot w_d(\overline{p},\overline{q}) \cdot w_c(\widetilde{I}_p, \widetilde{I}_q) \tag{4}$$

$$= \frac{\sum_{\overline{q}\in\Omega, D_{\overline{q}}\neq 0} D_{\overline{q}} \cdot \exp\left(-\frac{\|\overline{p}-\overline{q}\|}{2\delta_d^2}\right) \cdot \exp\left(-\frac{\left\|\widetilde{I}_p-\widetilde{I}_q\right\|}{2\delta_c^2}\right)}{\sum_{\overline{q}\in\Omega, D_{\overline{q}}\neq 0} \exp\left(-\frac{\|\overline{p}-\overline{q}\|}{2\delta_d^2}\right) \cdot \exp\left(-\frac{\left\|\widetilde{I}_p-\widetilde{I}_q\right\|}{2\delta_c^2}\right)}$$

and convert this expression into coordinate form:

$$\widetilde{D}(x,y) = \frac{\sum_{i,j} D(\overline{x}+i, \overline{y}+j) \cdot \exp\left(-\frac{i^2+j^2}{2\delta_d^2}\right) \cdot \exp\left(-\frac{(\widetilde{I}(x,y)-\widetilde{I}(x+i,y+j))^2}{2\delta_c^2}\right)}{\sum_{i,j} \exp\left(-\frac{i^2+j^2}{2\delta_d^2}\right) \cdot \exp\left(-\frac{(\widetilde{I}(x,y)-\widetilde{I}(x+i,y+j))^2}{2\delta_c^2}\right)}. \tag{5}$$

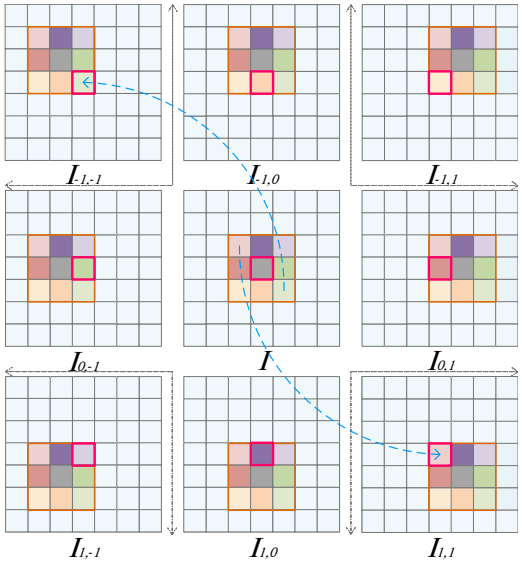


Figure 5: An example of the parallel matrix computations through image shifting under a filtering window size of 3×3.

As it is a 2× super-resolution in each layer, a small and fixed filtering window can be uniformly applied. Take a 3×3 filtering window as an example, by shifting the image up, down, left, and right as Figure 5 shows, we simultaneously interpolate the whole image using

$$\widetilde{D} = \frac{1}{W} \sum_{i,j} D_{i,j} \cdot W_d(i,j) \cdot W_c(\widetilde{I}, \widetilde{I}_{i,j}). \tag{6}$$

Here, (x, y) and $(\overline{x}, \overline{y})$ are horizontal and vertical coordinates of pixel p and $\overline{p}$, respectively, i and j are the offsets along horizontal and vertical directions, respectively, W_d denotes the spatial Gaussian kernel of the whole image which is only related to the offset displacement, W_c denotes the color Gaussian kernel of the whole image. The program is greatly accelerated by converting pixel-by-pixel calculations into parallel matrix computations.

4 CLASSIFICATION-BASED HIERARCHICAL DEPTH-MAP SUPER-RESOLUTION

For filter based interpolation methods, it is difficult to determine the reliable/valid points, that is, points should be weighted more heavily in the interpolation, especially at edge areas with complex colors. This is another reason why blurring and confusion are more likely to occur at edges as shown in Figure 1(b). In our C-HDS method, a context-adaptive classification strategy is introduced to reduce the contributions of invalid neighbors to the interpolation. Then, the interpolation is achieved via a joint trilateral filtering which includes the classification result.

The proposed context-adaptive classification strategy first classifies the RGB points under the supervision of depth, which means, pixels within different depth ranges will be classified into different classes. Therefore, pixels in the same class as the blank ones are regarded as reliable points, and are encouraged to make greater contributions than unreliable ones to the interpolation. The classification result not only has a linear correlation with the depth map, but also makes pixels in a same class more reliable to each

other. Figure 1(c) provides a visual result of the context adaptive classification based interpolation.

4.1 Depth-Guided Classification

Figure 6 shows an overview of the proposed context-adaptive classification. Firstly, the neighboring non-blank pixels of a to be interpolated point are classified into different classes according to the depth. Let $S_p = Sort\{D_{\overline{q}}\}$ denote the ascending order of $\{D_{\overline{q}}\}$, depth sequence of the neighboring non-blank points (shown in the 'to be interpolated depth map' in Figure 6). If the difference between two adjacent elements in S_p is larger than a threshold τ_d, i.e.

$$S_p(i+1) - S_p(i) > \tau_d, \tag{7}$$

the two elements are classified into different classes, as shown in Figure 6(a).

Then, with the aid of the RGB image, probabilities of a current point belongs to each class are calculated according to the Gaussian weights from the corresponding RGB points (shown in 'RGB image' in Figure 6) [12]:

$$M(p,n) = \frac{\sum_{L_q=n} w_c(\widetilde{I}_p, \widetilde{I}_q)}{\sum_q w_c(\widetilde{I}_p, \widetilde{I}_q)} \quad \text{for} \quad 1 \le n \le N, \tag{8}$$

where L_q denotes the class label of neighbor q, and N presents the total number of classes. If the maximum probability of p is larger than a threshold, candidates in the class holding this maximum probability are considered reliable, the current point is assigned to this class [2, 18]:

$$\widehat{n}_p = \begin{cases} n & \text{if } \max(M(p,n)) > \tau_m, 1 \le n \le N \\ 0 & \text{otherwise,} \end{cases} \tag{9}$$

where τ_m denotes the probability threshold. Figure 6(b) visualizes this step.

4.2 Joint Trilateral Filtering

To increase the weight of reliable points to the interpolation, we construct a joint trilateral filter by combining the classification result with the high-resolution RGB image and the to-be-upsampled depth map as follows:

$$\widetilde{D}_p = \frac{1}{W_p} \sum_{\overline{q}\in\Omega, D_{\overline{q}}\neq 0} D_{\overline{q}} \cdot w_d(\overline{p},\overline{q}) \cdot w_c(\widetilde{I}_p, \widetilde{I}_q) \cdot w_l(\widehat{n}_p, L_q), \tag{10}$$

where w_l denotes a piece-wise function according to the classification result $\widehat{n}$:

$$w_l(\widehat{n}_p, L_q) = \begin{cases} \alpha & \text{if } \widehat{n}_p \neq \emptyset, L_q = \widehat{n}_p \\ 1-\alpha & \text{if } \widehat{n}_p \neq \emptyset, L_q \neq \widehat{n}_p \\ 0.5 & \text{if } \widehat{n}_p = \emptyset, \end{cases} \tag{11}$$

where α is a preset constant. Finally, the current missing point p can be interpolated with this classification based joint trilateral filter as visualized in Figure 6(c).

With the proposed classification-based joint trilateral filter, contribution of invalid neighbors can be effectively reduced, so that edges of the upsampled depth map are well preserved. Note that the high-resolution RGB image in this upgraded method can also be replaced by the convolutional feature maps to construct a feature based trilateral filter in Equation 12. With the feature used

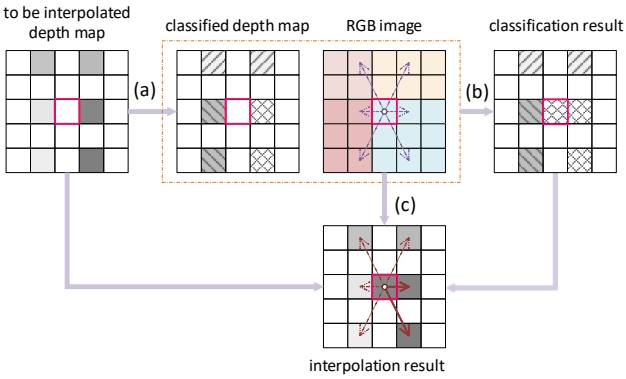


Figure 6: Classification based joint trilateral filtering. Three steps are: (a) classify the neighboring non-blank points according to depth values; (b) context-adaptive classification of the current blank point; and (c) joint trilateral interpolation of the current point. Contributions of the invalid neighbors to the interpolation are reduced as marked in lighter thin lines.

and classification based hierarchical depth map super-resolution (FC-HDS), better visual results can be obtained.

$$\widetilde{D}_p = \frac{1}{W_p} \sum_{\overline{q}\in\Omega, D_{\overline{q}}\neq 0} D_{\overline{q}} \cdot w_d(\overline{p},\overline{q}) \cdot w_c(\widetilde{F}_p^k, \widetilde{F}_q^k) \cdot w_l(\widehat{n}_p, L_q), \tag{12}$$

5 EXPERIMENTS

In this section, the experimental setup is introduced in the first place. Then in Section 5.1, we extensively compare our four methods, i.e. the basic HDS, F-HDS, C-HDS and FC-HDS with state-of-the-art approaches. The parameter analysis is presented in Section 5.2.

Datasets: Middlebury 2005 and 2006 contain 9 and 21 datasets, respectively, which are obtained using the technique of [35, 36]. Middlebury 2014 contains 33 datasets obtained using the technique of Scharstein et al. [34]. In our experiments, 6 open datasets in Middlebury 2005, all datasets in Middlebury 2006 and 23 open datasets in Middlebury 2014 are used. Both a high-resolution RGB image and its corresponding high-resolution depth map are included in each dataset. Image resolution of the Middlebury 2005 and 2006 datasets is approx. 1300×1100 pixels, while that of the Middlebury 2014 is approx. 3000×2000 pixels.

Implementation: Our system was built using Matlab on a PC with Core i7 2.9 GHz CPU and 32 GB RAM. In terms of datasets preprocessing, the high-resolution depth maps for these techniques are first inpainted (using SBU method) to remove the occlusions as shown in Figure 1, as some upsampling methods like JGU are only applicable to low-resolution depth maps with no occlusions [32]. Then, to generate the low-resolution depth, we downsample the inpainted image in 2×, 4×, 8×, 16×, respectively. To the best of our knowledge, 16× super-resolution has rarely been discussed in previous work.

Evaluation Protocol: We use both mean squared error (MSE) and bad pixel rate error (BPR) for quantitative evaluations. 'Bad' pixels are those whose value deviates from the ground truth by more than

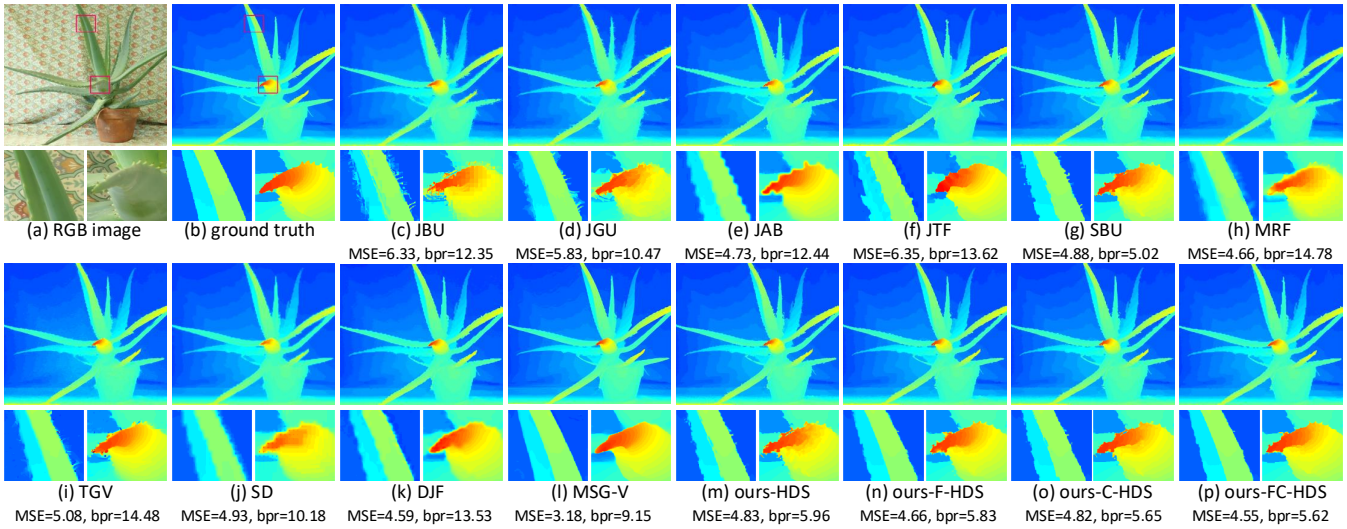


Figure 7: Comparison of 8× super-resolution result of *Aloe*.

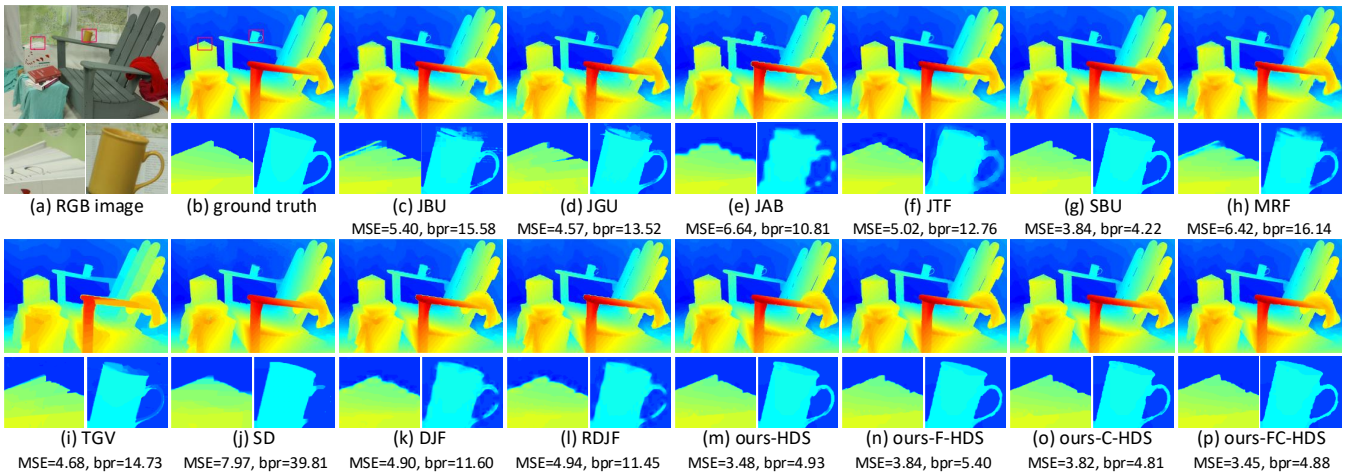


Figure 8: Comparison of 16× super-resolution result of *Adirondack*.

one disparity level. MSE has been widely used to rank the quality of upsampled depth maps, however, BPR is less used. In practice, even with high MSE score, the quality of upsampled depth maps will not be convincing if BPR measure is low. Such case is mostly caused by over-smoothing. Thus we consider BPR to be particularly important in our evaluation of depth maps. We also use MSE_disc and BPR_disc to evaluate the performance near discontinuities [32].

5.1 Visual, Quantitative and Runtime Evaluations

Figure 7 shows visual results and comparison of 8× super-resolution on the *Aloe* dataset [35], and Figure 8 shows 16× super-resolution on the *Adirondack* dataset [34]. We present quantitative evaluations in Tables 1 and 2[1], and runtime measurements in Table 3. In these tables, we compare the proposed methods with state-of-the-art interpolation-, optimization- and learning-based methods.

[1]The results are separated into two tables as MSG-V can only be executed on datasets Middlebury 2005 and 2006, and would run into an error on Middlebury 2014, while the other methods can be executed on all 3 datasets.

Comparison to Interpolation-based Methods: As mentioned, with interpolation based methods, missing depths are filled with a mixture of their non-blank neighbors. This may lead to several issues: (1) the non-blank neighbors may be distant, especially under a high super-resolution scale. This further contributes to depth confusions as shown in Figure 7(c) and Figure 8(c); (2) as it is difficult to determine which neighbors are suitable for interpolation, a global search is proposed in JGU [24] to seek the most closest neighbors in geodesic distance. However, depth bleeding is brought due to the recursive search, as shown in Figure 7(d) and Figure 8(d). Both issues bring extra errors into quantitative evaluation, both for the global result and near discontinuities. The situation worsens as the super-resolution scale increases. These two issues can be resolved through a hierarchical method with a super-resolution scale of 2 for each layer, so that non-blank points can be searched close to the missing points. Our classification addresses this by selecting suitable neighbors for interpolation; (3) some methods, like JAB [19] and JTF [25], generate the final super-resolution result by refining depth boundaries of a simple initial interpolation result, which may

Table 1: Quantitative evaluation for different super-resolution scales on all the three datasets Middlebury 2005, 2006 and 2014.Grouping of methods, from top to bottom: filter-based, optimization-based, learning-based, ours.

Methods	MSE				MSE_disc				BPR(%)				BPR_disc(%)			
	2×	4×	8×	16×	2×	4×	8×	16×	2×	4×	8×	16×	2×	4×	8×	16×
JBU [21]	2.63	3.67	4.91	6.66	10.98	14.67	17.39	19.94	1.12	2.63	6.56	16.08	20.49	34.88	44.80	54.66
JGU [24]	2.61	3.48	4.55	6.19	11.98	14.79	16.80	19.13	1.50	2.79	5.79	12.82	26.30	34.14	41.53	51.09
JAB [19]	2.44	3.16	4.38	6.44	11.44	13.85	17.00	21.55	2.10	3.76	6.90	13.62	39.42	49.66	62.47	73.48
JTF [25]	2.33	4.54	4.77	5.93	10.92	18.89	17.69	18.88	2.18	3.54	7.10	14.86	42.83	48.80	63.95	78.70
SBU [32]	2.51	3.29	4.19	5.40	10.29	13.15	15.35	17.28	0.65	1.24	**2.72**	6.57	12.04	19.26	**26.63**	**33.39**
MRF [7]	2.62	3.29	4.49	6.56	10.93	13.16	16.24	20.39	2.05	3.31	7.04	16.53	39.02	47.13	55.45	63.40
TGV [9]	2.57	3.62	4.86	8.47	10.56	14.00	16.64	20.95	1.61	3.61	7.29	21.34	21.54	36.54	45.17	57.82
SD [13]	2.41	3.26	4.44	7.18	**9.72**	12.78	15.96	20.28	1.52	3.15	6.70	29.12	26.09	40.66	55.15	71.39
DJF [23]	—	2.78	3.79	5.42	—	**11.60**	14.35	17.34	—	4.41	7.41	20.33	—	51.66	65.29	74.98
RDJF [23]	—	**2.72**	**3.74**	5.32	—	11.68	**14.30**	17.10	—	3.12	5.97	13.78	—	46.54	59.06	74.58
Ours-HDS	**2.24**	3.14	4.00	**5.17**	10.48	13.50	15.11	**16.74**	**0.53**	1.30	3.10	6.99	10.59	20.63	30.25	39.04
Ours-F-HDS	2.28	3.26	4.09	5.21	10.78	14.13	15.55	16.99	0.69	1.55	3.31	6.92	13.89	24.86	35.08	41.13
Ours-C-HDS	2.52	3.35	4.17	5.38	10.77	13.65	15.48	17.43	0.54	**1.22**	2.89	6.59	**10.22**	**18.84**	27.12	35.22
Ours-FC-HDS	2.37	3.32	4.20	5.38	11.22	14.37	16.04	17.75	0.57	1.41	3.04	**6.50**	11.28	22.15	31.29	37.68

Table 2: Quantitative evaluation between our methods and MSG-V based on Middlebury 2005 and 2006 only, as 2014 dataset caused a 'out of memory' error for MSG-V.

Methods	4×				8×			
	MSE	MSE_disc	BPR	BPR_disc	MSE	MSE_disc	BPR	BPR_disc
MSG-V [42]	**1.07**	**4.55**	1.74	29.38	**1.77**	**6.83**	3.76	43.64
Ours-HDS	2.29	10.26	1.20	19.48	2.88	11.42	3.25	30.12
Ours-F-HDS	2.45	11.07	1.43	23.83	2.99	11.94	3.26	34.59
Ours-C-HDS	2.45	10.32	**1.16**	**17.93**	2.99	11.60	3.09	**27.35**
Ours-FC-HDS	2.47	11.15	1.33	21.58	3.03	12.15	**3.05**	31.18

Table 3: Runtimes averaged on the Middlebury 2006/2014 datasets.

Methods	Runtime (s)			
	2×	4×	8×	16×
JBU [21]	0.83/4.04	0.99/5.01	1.04/5.25	1.02/5.31
JGU [24]	62.15/213.82	123.83/446.85	192.09/814.27	260.24/1254.10
JAB [19]	11.63/36.36	34.58/79.87	64.41/202.20	299.21/444.65
JTF [25]	78.38/156.88	167.66/250.99	285.57/463.32	500.48/871.69
SBU [32]	294.65/450.77	323.83/465.25	344.65/485.31	349.87/492.28
MRF [7]	117.08/609.93	113.73/595.57	106.85/565.05	101.45/560.72
TGV [9]	322.33/1137.80	531.02/1816.62	743.06/2702.28	983.13/3641.70
SD [13]	14.87/61.94	20.77/95.33	44.42/204.15	91.36/418.21
DJF [23]	—	6.07/218.94	5.87/221.49	5.81/205.29
RDJF [23]	—	5.90/306.61	5.79/231.81	9.26/760.94
MSG-V [42]	—	60.90/—	66.91/—	—/—
Ours-HDS	0.25/1.12	0.30/1.45	0.32/1.62	0.34/1.64
Ours-F-HDS	1.40/6.11	1.74/7.98	1.93/9.15	2.11/10.12
Ours-C-HDS	1.21/7.32	1.07/5.29	1.03/5.60	1.15/5.71
Ours-FC-HDS	1.86/7.99	2.39/10.83	2.80/12.62	3.11/13.86

produce varying degrees of depth distortions (shown in (e) and (f) of Figures 7 and 8). Due to simple initial interpolation, which tends to over-smoothing, MSE and MSE_disc maintain a normal level. Nevertheless, BPR and BPR_disc go extremely high.

The SBU method [32] solves the above problems to a large extent based on the idea of segmentation. However, in boundary regions with rich textures, the segmentation result can be inaccurate, which may further degrade the final super-resolution result, as Figure 7(g) and Figure 8(g) show. Besides, compared with the global segmentation, the proposed local context-adaptive classification can better detect edges to reduce serious artifacts in these regions.

In terms of runtime, our basic HDS provides strong performance by eliminating the need to search suitable non-blank neighbors and the addition of parallel matrix computation.

Comparison to Optimization-based Approaches: The optimization-based methods MRF [7], TGV [9] and SD [13] demonstrate over-smoothing in certain areas in order to solve the aliasing artifacts caused by simple initial interpolation, as (h), (i) and (j) of Figures 7 and 8 show. This can also cause loss of detail in images.

The optimization-based methods seem to perform well on MSE and MSE_disc, but this is actually caused by over-smoothing, which on the other hand can lead to high BPR and BPR_disc. When the situation is serious enough or the super-resolution scale is large enough that the details begin to disappear, all MSE, MSE_disc, BPR and BPR_disc begin to increase sharply. This is also why optimization-based methods are the most unstable. In addition, optimization is also computationally expensive (see Table 3).

Comparison to Learning-based Methods: Learning-based methods such as DJF and RDJF [23] also yield the final super-resolution result by iteratively optimizing the depth map. But unlike traditional optimization based methods, the filters are obtained through training with a loss function positively correlated with MSE. Therefore, when the super-resolution scale is small, MSE and MSE_disc are low. Even for larger super-resolution factors, MSE and MSE_disc are controlled to rise to some extent. However, BPR and BPR_disc are always high. As mentioned before, low MSE and high BPR leads the learning-based methods loss of detail less of an issue, as shown in (k) and (l) of Figures 7 and 8.

By introducing a number of perceptual metrics as loss, MSG-V[42] gives the best accuracy across previous deep learning methods. Note that we present only 4× and 8× super-resolution results of MSG-V in Table 2 as the 16× MSG-V model is not available. Besides, the implementation leads to a 'out of memory' issue on a 32G host when running on Middlebury 2014 dataset, thus the evaluations in Table 2 are based only on Middlebury 2005 and 2006 datasets. We observe that MSG-V gives the best MSE measure but lower BPR score, especially in the discontinuous regions, meaning their resulting depth at boundary regions is over-smoothed where the details would be lost (see Figure 7(l)).

In our improved method, the learned context-aware maps are utilized to build the feature based filters, which contribute a lot to improve the visual effect of the output high-resolution depth maps, as (n) and (p) of Figures 7 and 8 shows. In addition, our program runs fast while the runtimes of deep learning methods are susceptible to image resolution, and increase sharply with increasing resolution.

In summary, the proposed methods, comparing to the previous interpolation- and optimization-based methods, generate higher quality upsampled depth maps in both visual and quantitative evaluations, especially for the large scale case (16× scale). The learning-based methods, i.e. DJF and RDJF in Table 1 and MSG-V in Table 2, have slightly better MSE values when the up-sampling rate is not very high (see 4× and 8×), the price is the over-smoothing during the up-sampling. Our methods effectively avoid the over-smoothing and obtain clearer and more accurate boundaries, and further lead to better BRP values, especially on the boundaries (see BRP and BRP_disc). Moreover, C-HDS particularly shows the best BPR measurement with fairly low MSE values guaranteed for most of the trials. F-HDS provides visually best results due to the use of extracted shallow and multi-layer feature maps. FC-HDS takes advantage of C-HDS and F-HDS, and yields the best results by taking both visual comparison and quantitative analysis into account. In addition, the proposed methods outperform most of the baselines in terms of runtime, it is worth mentioning that HDS achieves 7.6Hz using parallel computation on CPU.

5.2 Parameter Analysis

The depth threshold τ_d is used to determine if classification will take place under the current sliding window. If the depth gap between any two adjacent depths is bigger than τ_d, classification will be conducted before the interpolation. Otherwise if all depth differences under the current sliding window are small enough, i.e. neighbors are within similar depth ranges, classification will no longer be necessary. We set $\tau_d = 5$ in our experiments empirically.

The membership degree threshold τ_m is used to control the assignment of a current point. If the maximum membership degree is larger than τ_m, the current point will be assigned to the class with the maximum membership degree. Otherwise, the assignment will not proceed and all neighbors will be treated equally. As shown in Figure 9, the larger the τ_m, the fewer points will be assigned, the smoother the image will be, the smaller the MSE and the larger the BPR. τ_m is set to 0.8 in our experiments.

We use the parameter α to control the contributions of neighbors to the interpolation. Assuming that the current point is assigned to a class, pixels in this class should contribute more than pixels in other classes during interpolation. Theoretically, the larger α, the sharper the image will be, the larger the MSE and the smaller the BPR. In addition, α usually has a greater impact on BPR than MSE. However, as α tends to one, $(1-\alpha)$ tends towards zero, which means all the contributions of points in other classes to the interpolation will be cut off referring to Equation 10. This will further cause a rise in both MSE and BPR values, and MSE rises sharply due to its sensitivity. We generally set α close to 1, 0.9 in our experiments.

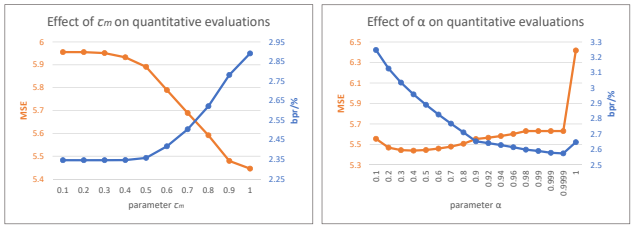


Figure 9: Effects of parameters on quantitative evaluations on Middlebury 2006. Left: effect of τ_m on MSE and BPR evaluations. Right: effect of α on MSE and BPR evaluations.

6 CONCLUSIONS

A fast and high quality HDS method is proposed in this work. Instead of one-step upsampling, a hierarchical image pyramid strategy is adopted, that is, we upsample the low-resolution depth map with a sampling scale of 2 at each and every layer, under the guidance of the pre-downsampled RGB image with a same resolution in the same layer. To obtain more sharp and clear depth edges, we construct a context-adaptive classification based trilateral filter to upgrade the basic HDS method to a C-HDS method. Given the original images with the same quality, both the proposed basic HDS and the upgraded C-HDS outperform the current state-of-the-art approaches, especially in the case of large scale (16×). And the higher quality of original depth maps will result in higher up-sampling quality. In addition, the program is stable, training-free and easy to implement, with run times that exceed other methods to the best of our knowledge based on claimed runtimes.

Beyond super-resolution, the proposed method is also applicable to depth map inpainting. Specifically, blank pixels will be eroded away during the degradation of the depth map, and be filled in during the upsampling process. In addition, due to the strong interpretability, our methods can be simply and widely used for other types of fusion data processing which are similar to RGB-D data , such as RGB-T (thermal) data. Like most methods, our method is insensitive to thin lines, which are easy to lose in low resolution depth maps, and difficult to be retrieved during the upsampling. Especially in the case of complete loss of depth ranges, completion seems to be impossible. Our future work is to find the linear correspondence between RGB images and depth maps by using a depth supervised RGB image pixel-level classification strategy. With this, the low-resolution depth map can be upsampled with some guided filters under the guidance of the classification result [15, 16].

ACKNOWLEDGMENTS

The authors are grateful for the freely distributed Middlebury stereo datasets [34–36]. This work was supported by the EPSRC grant CAMERA EP/M023281/1 and EP/T022523/1.

REFERENCES

[1] Sari Awwad, Fairouz Hussein, and Massimo Piccardi. 2015. Local Depth Patterns for Tracking in Depth Videos. In *International Conference on Multimedia*. Association for Computing Machinery, New York, NY, USA, 1115–1118. https://doi.org/10.1145/2733373.2806295

[2] A. M. Bensaid, L. O. Hall, J. C. Bezdek, L. P. Clarke, M. L. Silbiger, J. A. Arrington, and R. F. Murtagh. 1996. Validity-guided (re)clustering with applications to image segmentation. *IEEE Transactions on Fuzzy Systems* 4, 2 (May 1996), 112–123. https://doi.org/10.1109/91.493905

[3] Derek Chan, Hylke Buisman, Christian Theobalt, and Sebastian Thrun. 2008. A Noise-aware Filter for Real-time Depth Upsampling. In *ECCV Workshops*. IEEE, Marseille, France.

[4] K. N. Chaudhury. 2013. Acceleration of the Shiftable $O(1)$ Algorithm for Bilateral Filtering and Nonlocal Means. *IEEE Transactions on Image Processing* 22, 4 (April 2013), 1291–1300. https://doi.org/10.1109/TIP.2012.2222903

[5] Ruijin Chen and Wei Gao. 2020. Color-Guided Depth Map Super-Resolution Using a Dual-Branch Multi-Scale Residual Network with Channel Interaction. *Sensors* 20 (2020), 6. https://doi.org/10.3390/s20061560

[6] Longquan Dai, Mengke Yuan, and Xiaopeng Zhang. 2016. Speeding up the bilateral filter: A joint acceleration way. *IEEE Transactions on Image Processing* 25, 6 (2016), 2657–2672.

[7] James Diebel and Sebastian Thrun. 2006. An Application of Markov Random Fields to Range Sensing. In *NIPS*. MIT Press, Vancouver, B.C., Canada, 291–298. http://papers.nips.cc/paper/2837-an-application-of-markov-random-fields-to-range-sensing.pdf

[8] Elmar Eisemann and Frédo Durand. 2004. Flash photography enhancement via intrinsic relighting. *ACM Transactions on Graphics (Proceedings of SIGGRAPH)* 23, 3 (August 2004), 673–678. https://doi.org/10.1145/1186562.1015778

[9] David Ferstl, Christian Reinbacher, Rene Ranftl, Matthias Rüther, and Horst Bischof. 2013. Image Guided Depth Upsampling using Anisotropic Total Generalized Variation. In *ICCV*. IEEE Computer Society, USA, 993–1000.

[10] Sergi Foix, Guillem Alenyà, and Carme Torras. 2011. Lock-in Time-of-Flight (ToF) Cameras: A Survey. *IEEE Sensors Journal* 11 (2011), 1917–1926.

[11] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. 2014. Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation. In *CVPR*. IEEE Computer Society, USA, 580–587. https://doi.org/10.1109/CVPR.2014.81

[12] Jing Gu, Licheng Jiao, Shuyuan Yang, and Fang Liu. 2018. Fuzzy Double C-Means Clustering Based on Sparse Self-Representation. *IEEE Transactions on Fuzzy Systems* 26, 2 (April 2018), 612–626. https://doi.org/10.1109/TFUZZ.2017.2686804

[13] Bumsub Ham, Minsu Cho, and Jean Ponce. 2018. Robust Guided Image Filtering Using Nonconvex Potentials. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 40, 1 (Jan 2018), 192–207. https://doi.org/10.1109/TPAMI.2017.2669034

[14] Jungong Han, Ling Shao, Dong Xu, and Jamie Shotton. 2013. Enhanced Computer Vision With Microsoft Kinect Sensor: A Review. *IEEE Transactions on Cybernetics* 43, 5 (oct 2013), 1318–1334. https://doi.org/10.1109/tcyb.2013.2265378

[15] Kaiming He, Jian Sun, and Xiaoou Tang. 2013. Guided Image Filtering. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 35, 6 (June 2013), 1397–1409. https://doi.org/10.1109/TPAMI.2012.213

[16] Qibin Hou, Ming-Ming Cheng, Xiaowei Hu, Ali Borji, Zhuowen Tu, and Philip H. S. Torr. 2019. Deeply Supervised Salient Object Detection with Short Connections. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 41, 4 (April 2019), 815–828. https://doi.org/10.1109/TPAMI.2018.2815688

[17] Tak-Wai Hui, Chen Change Loy, and Xiaoou Tang. 2016. Depth map super-resolution by deep multi-scale guidance. In *European conference on computer vision*. Springer, Springer International Publishing, Cham, 353–369.

[18] J. M. Keller, M. R. Gray, and J. A. Givens. 1985. A fuzzy K-nearest neighbor algorithm. *IEEE Transactions on Systems, Man, and Cybernetics* SMC-15, 4 (July 1985), 580–585. https://doi.org/10.1109/TSMC.1985.6313426

[19] Joohyeok Kim, Gwanggil Jeon, and Jechang Jeong. 2014. Joint-adaptive bilateral depth map upsampling. *Signal Processing: Image Communication* 29, 4 (2014), 506–513. https://doi.org/10.1016/j.image.2014.01.011

[20] Andreas Kolb, Erhardt Barth, Reinhard Koch, and Rasmus Larsen. 2010. Time-of-Flight Cameras in Computer Graphics. *Computer Graphics Forum* 29, 1 (March 2010), 141–159. https://doi.org/10.1111/j.1467-8659.2009.01583.x

[21] Johannes Kopf, Michael F. Cohen, Dani Lischinski, and Matt Uyttendaele. 2007. Joint bilateral upsampling. *ACM Transaction on Graphics* 26, 3 (July 2007), 96. https://doi.org/10.1145/1276377.1276497

[22] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. 2012. ImageNet Classification with Deep Convolutional Neural Networks. In *NIPS*. Association for Computing Machinery, New York, NY, USA, 84–90.

[23] Yijun Li, Jia-Bin Huang, Ahuja Narendra, and Ming-Hsuan Yang. 2016. Deep Joint Image Filtering. In *ECCV*. Springer International Publishing, Cham, 154–169.

[24] Ming-Yu Liu, Oncel Tuzel, and Yuichi Taguchi. 2013. Joint Geodesic Upsampling of Depth Images. In *CVPR*. IEEE Computer Society, Los Alamitos, CA, USA, 169–176.

[25] Kai-Han Lo, Yu-Chiang Frank Wang, and Kai-Lung Hua. 2018. Edge-Preserving Depth Map Upsampling by Joint Trilateral Filter. *IEEE Transactions on Cybernetics* 48 (2018), 371–384.

[26] Jonathan Long, Evan Shelhamer, and Trevor Darrell. 2015. Fully convolutional networks for semantic segmentation. In *CVPR*. IEEE Press, Boston, MA, USA, 3431–3440.

[27] Bruno Macchiavello, Camilo Dorea, Edson M. Hung, Gene Cheung, and Wai-Tian Tan. 2014. Loss-Resilient Coding of Texture and Depth for Free-Viewpoint Video Conferencing. *Trans. Multi.* 16, 3 (April 2014), 711–725. https://doi.org/10.1109/TMM.2014.2299768

[28] Ilya Makarov, Vladimir Aliev, and Olga Gerasimova. 2017. Semi-Dense Depth Interpolation Using Deep Convolutional Neural Networks. In *International Conference on Multimedia*. Association for Computing Machinery, New York, NY, USA, 1407–1415. https://doi.org/10.1145/3123266.3123360

[29] Patrick Ndjiki-Nya, Martin Köppel, Dimitar Doshkov, Haricharan Lakshman, Philipp Merkle, Karsten Müller, and Thomas Wiegand. 2010. Depth Image-Based Rendering With Advanced Texture Synthesis for 3-D Video. *IEEE Transactions on Multimedia* 13 (2010), 453–465.

[30] Simon Niklaus and Feng Liu. 2018. Context-Aware Synthesis for Video Frame Interpolation. In *CVPR*. IEEE Computer Society, Los Alamitos, CA, USA, 1701–1710.

[31] Georg Petschnigg, Richard Szeliski, Maneesh Agrawala, Michael Cohen, Hugues Hoppe, and Kentaro Toyama. 2004. Digital photography with flash and no-flash image pairs. *ACM Transactions on Graphics (Proceedings of SIGGRAPH)* 23, 3 (August 2004), 664–672. https://doi.org/10.1145/1186562.1015777

[32] Yiguo Qiao, Licheng Jiao, Shuyuan Yang, and Biao Hou. 2019. A Novel Segmentation Based Depth Map Up-Sampling. *IEEE Transactions on Multimedia* 21, 1 (Jan 2019), 1–14. https://doi.org/10.1109/TMM.2018.2845699

[33] Christian Richardt, Carsten Stoll, Neil A. Dodgson, Hans-Peter Seidel, and Christian Theobalt. 2012. Coherent Spatiotemporal Filtering, Upsampling and Rendering of RGBZ Videos. *Computer Graphics Forum (Proceedings of Eurographics)* 31, 2 (May 2012), 247–256. https://doi.org/10.1111/j.1467-8659.2012.03003.x

[34] Daniel Scharstein, Heiko Hirschmüller, York Kitajima, Greg Krathwohl, Nera Nešić, Xi Wang, and Porter Westling. 2014. High-Resolution Stereo Datasets with Subpixel-Accurate Ground Truth. In *German Conference on Pattern Recognition*. Springer International Publishing, Cham, 31–42.

[35] Daniel Scharstein and Chris Pal. 2007. Learning Conditional Random Fields for Stereo. In *CVPR*. IEEE Computer Society, Los Alamitos, CA, USA, 1–8. https://doi.org/10.1109/CVPR.2007.383191

[36] Daniel Scharstein and Richard Szeliski. 2003. High-accuracy stereo depth maps using structured light. In *CVPR*. IEEE Computer Society, USA, 195–202. https://doi.org/10.1109/CVPR.2003.1211354

[37] Jamie Shotton, Ross Girshick, Andrew Fitzgibbon, Toby Sharp, Mat Cook, Mark Finocchio, Richard Moore, Pushmeet Kohli, Antonio Criminisi, Alex Kipman, and Andrew Blake. 2013. Efficient Human Pose Estimation from Single Depth Images. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 35, 12 (December 2013), 2821–2840. https://doi.org/10.1109/TPAMI.2012.241

[38] Karen Simonyan and Andrew Zisserman. 2015. Very Deep Convolutional Networks for Large-Scale Image Recognition. In *ICLR*. IEEE, Kuala Lumpur, Malaysia, 730–734.

[39] Mashhour Solh and Ghassan Al-Regib. 2012. Hierarchical Hole-Filling For Depth-Based View Synthesis in FTV and 3D Video. *IEEE Journal of Selected Topics in Signal Processing* 6 (2012), 495–504.

[40] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. 2015. Going Deeper With Convolutions. In *CVPR*. IEEE, Los Alamitos, CA, USA, 1–9.

[41] Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. 2018. Deep image prior. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, Salt Lake City, UT, USA, 9446–9454.

[42] Oleg Voynov, Alexey Artemov, Vage Egiazarian, Alexander Notchenko, Gleb Bobrovskikh, Evgeny Burnaev, and Denis Zorin. 2019. Perceptual Deep Depth Super-Resolution. In *ICCV*. IEEE, Seoul, Korea (South), 5653–5663.

[43] Anran Wang, Jianfei Cai, Jiwen Lu, and Tat-Jen Cham. 2016. Modality and Component Aware Feature Fusion for RGB-D Scene Classification. In *CVPR*. IEEE, Las Vegas, NV, USA, 5995–6004. https://doi.org/10.1109/CVPR.2016.645

[44] Jiang Wang, Zicheng Liu, Ying Wu, and Junsong Yuan. 2012. Mining Actionlet Ensemble for Action Recognition with Depth Cameras. In *CVPR*. IEEE, Providence, RI, USA, 1290–1297.

[45] Yucheng Wang, Jian Zhang, Zicheng Liu, Qiang Wu, Philip A. Chou, Zhengyou Zhang, and Yunde Jia. 2016. Handling Occlusion and Large Displacement Through Improved RGB-D Scene Flow Estimation. *IEEE Transactions on Circuits and Systems for Video Technology* 26, 7 (July 2016), 1265–1278. https://doi.org/10.1109/TCSVT.2015.2462011

[46] Lu Xia and J.K. Aggarwal. 2013. Spatio-temporal Depth Cuboid Similarity Feature for Activity Recognition Using Depth Camera. In *CVPR*. IEEE, Portland, OR, USA, 2834–2841.

[47] Jingyu Yang, Xinchen Ye, Kun Li, Chunping Hou, and Yao Wang. 2014. Color-Guided Depth Recovery From RGB-D Data Using an Adaptive Autoregressive

Model. *IEEE Transactions on Image Processing* 23, 8 (Aug 2014), 3443–3458. https://doi.org/10.1109/TIP.2014.2329776

[48] Zhengyou Zhang. 2012. Microsoft Kinect Sensor and Its Effect. *IEEE MultiMedia* 19, 2 (February 2012), 4–10. https://doi.org/10.1109/MMUL.2012.24

[49] Lijun Zhao, Huihui Bai, Jie Liang, Anhong Wang, Bing Zeng, and Yao Zhao. 2019. Local activity-driven structural-preserving filtering for noise removal and image smoothing. *Signal Processing* 157 (2019), 62 – 72. https://doi.org/10.1016/j.sigpro.2018.11.012

[50] Lijun Zhao, Huihui Bai, Jie Liang, Bing Zeng, Anhong Wang, and Yao Zhao. 2019. Simultaneous color-depth super-resolution with conditional generative adversarial networks. *Pattern Recognition* 88 (2019), 356 – 369. https://doi.org/10.1016/j.patcog.2018.11.028